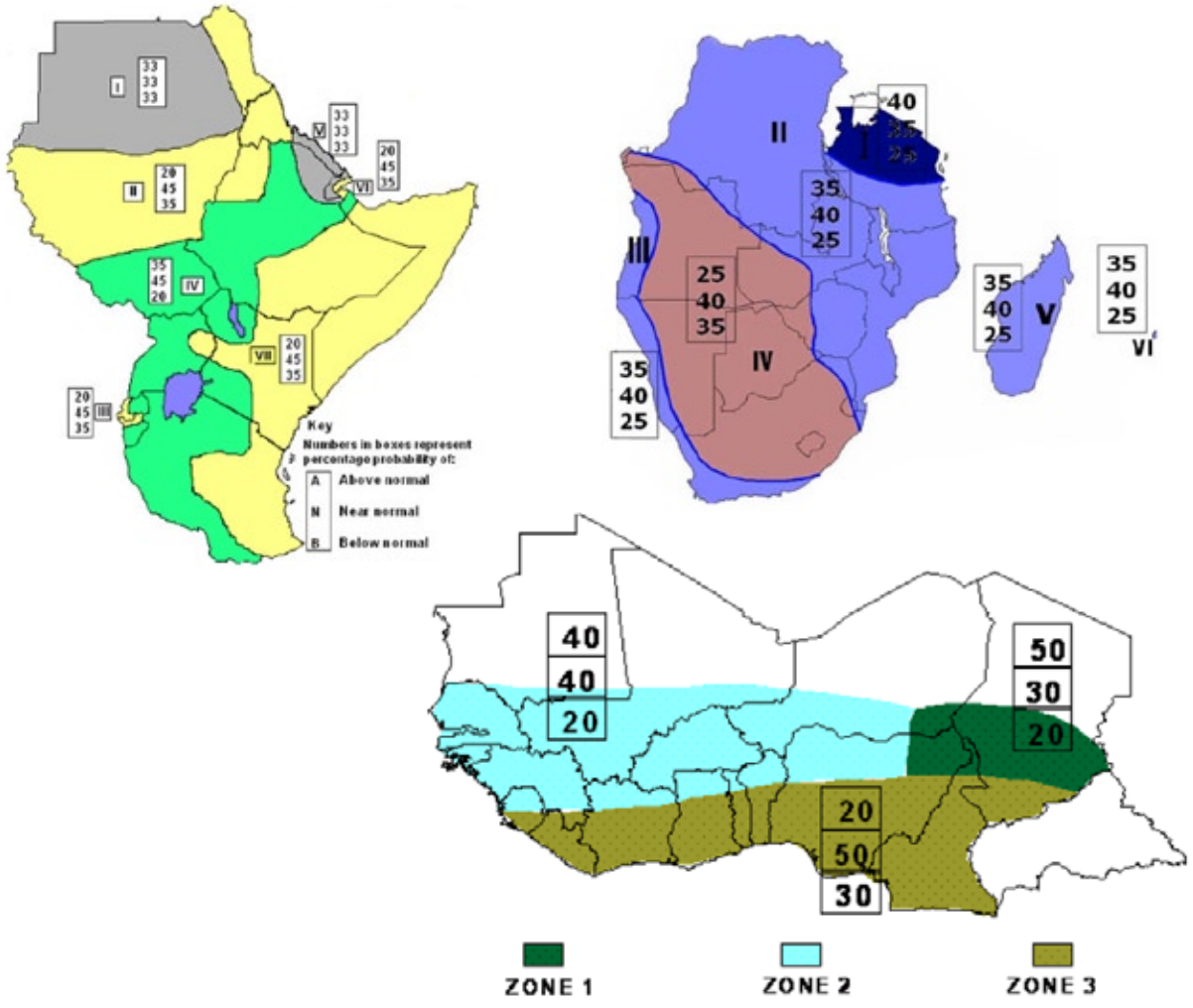
Technical Report #09-02
RCOF Review 2008

Position Paper: **Verification of African RCOF Forecasts**

by Dr. Simon Mason and Mr. Simbarashe Chidzambwa



Dr Simon J. Mason
International Research Institute for Climate and Society
The Earth Institute of Columbia University
61 Route 9W Palisades, NY 10964
simon@iri.columbia.edu

Mr Simbarashe Chidzambwa
African Centre of Meteorological Application for Development
85 Avenue des Ministères
BP 13184
Niamey, Niger

EXECUTIVE SUMMARY

Now that the Regional Climate Outlook Forums in Africa have been operating for over ten years, an evaluation of the skill of these forecasts is possible. For most other regions in which RCOFs have been held there are fewer forecasts available for any detailed diagnostic verification, but many of the lessons learnt from a verification of the Africa RCOF forecasts are relevant globally. In addition, the identification of appropriate verification procedures has relevance globally, since forecasts are presented in similar formats at all the RCOFs.

Forecasts are verified from three RCOFs in Africa: for Southern Africa forecasts are verified for the October – December (early-season) and January – March (late-season) summer rainfall season; for the Greater Horn, the target seasons are March – May and September – December; and for West Africa forecasts are for July – September. All three regions indicate some evidence of positive skill, meaning that they contain useful information that could potentially have been used to achieve some form of benefit. In addition to the numerous other benefits, such as the development of capacity within the climate services of the National Meteorological Services in most of Africa, the positive skill provides a powerful endorsement to the RCOF process. However, the forecasts do show clear evidence of systematic errors, and in some cases the positive skill may not be immediately apparent to users. There is thus considerable scope for improvement.

The most ubiquitous error is for the forecasters to hedge the forecasts towards high probabilities on the normal category. The probabilities for the normal category are therefore consistently higher than they should be, and the normal rainfall occurred notably much less frequently and extensively than implied by the forecasts. This hedging is an effect of an ongoing deterministic interpretation of the forecasts and the wish to avoid the risk of the forecasts being interpreted as in error by two categories (which is possible only if below- or above-normal rainfall is forecast). In addition to this over-forecasting of the normal category, there is little or no evidence of any skill in forecasting increased probabilities for this category. More generally, the probabilities for all categories typically show poor reliability, and there is a need to implement improved procedures for defining the probabilities. In most cases the poor reliability reflects over-confidence (increases and decreases in probabilities are too large), which points to a need to review the scientific bases for some of the predictions.

Over the approximately 10-year verification period, below-normal rainfall was predominant in the Greater Horn in both seasons, in West Africa for the July – September period, and in Southern Africa for January – March. The RCOFs did not provide any clear indications of these trends, which has to be acknowledged as a notable failure. Again, the need for a serious review of the scientific bases for how the forecasts are currently made needs to be undertaken, and an assessment of the potential benefit of making greater use of Global Producing Centre products should be conducted.

Apart from these considerations of the skill of the forecasts, ambiguities in the precise meaning of the forecasts occur because of the way in which the forecasts are constructed. Specifically it is not clear whether the forecasts are meant to be interpreted only as regional averages, and, if so, what precisely the regions are over which the averages should be calculated. It is recommended that this ambiguity be addressed by careful consideration of how the forecasts are constructed; specifically, greater consistency is required in the ways in which the forecasts are made for each country before the consensus building step.

1. Introduction

Since the first such meeting in September 1997, countries have come together each year in various regional groupings to issue a consensus seasonal climate forecast for their respective regions. The broad benefits of these so called Regional Climate Outlook Forums (RCOFs) in terms of forecast generation, dissemination, and capacity building are widely recognized such that the RCOF process is becoming replicated in various ways around the globe (Ogallo *et al.* 2008). Seasonal forecasts are meant to benefit decision making through their application in various sectors such as agriculture, water resources, energy, health, manufacturing, etc.. However, the potential usefulness of seasonal forecasts has not been fully exploited for a number of reasons including: limited access to forecasts due to communication problems; issuance of forecasts in forms not applicable to particular sectors; poor interpretation due to little or no understanding of the forecasts; and lack of information about the quality of the forecast which ultimately leads to a lack of confidence in the information.

Since the commencement of the RCOFs, no comprehensive evaluation of the quality of the forecasts has been assessed, except for a preliminary assessment for the forecasts from South-eastern South America has been conducted (Berri *et al.* 2005). There are a number of reasons why seasonal forecasts should be evaluated (Jolliffe and Stephenson 2003), and there are numerous methods available to evaluate them. Knowledge about the quality of the seasonal forecasts would help users to understand the risks and uncertainties involved when considering the information. Evaluating the forecast not only benefits the users but the producers of the forecast by way of identifying weaknesses and improving on them. There are two main objectives in this white paper: to provide simple indications to existing and potential users about the quality of the forecasts that have been issued by the RCOFs; and to provide diagnostics to the forecasters themselves so that forecasts can be improved. The focus of this verification analysis is exclusively on the African RCOFs, primarily because they have a relatively long, and continuous history, but the objective is to spawn similarly comprehensive verification analyses for the RCOFs in other parts of the world, and to encourage more detailed analyses of the African RCOFs using improved datasets and the progressively expanding history of forecasts.

Further details on the reasons for this verification exercise and the associated methods applied are expounded in section 3. The results and discussions are provided in section 4 with a final summary in section 5. First, however, further details on the RCOF process and the probabilistic forecasts are given in section 2, where recommendations are made on how to verify prepare the data for verification analysis.

2. RCOF Forecasts

a. The Computation of RCOF Forecasts

To date, seasonal forecasting in the RCOF process has been based primarily on the formulation of multiple linear regression equation models using the classical statistical forecasting methods based on relationships between historical observations of predictors and predictands. A set of identified input predictor values (SSTs) observed or expected before the time that the forecast is issued are input into regression model to forecast the future values of the predictands (rainfall). The forecast time lag is therefore built into the regression model (Wilks 2006). In more recent years, regression-based models have been used to downscale the outputs of general circulation models (GCMs) from some of the WMO-designated Global Producing Centres (GPCs) and other dynamical modelling centres.

However, the SST-based models are usually given most weight in the development of the consensus forecast.

Regression models are used to make predictions at national or sub-national scale. These are then collated, and regions that have the same expected climate variability are grouped together and the probabilities are estimated based on methods including contingency tables, and ensemble probabilities from general circulation model outputs (GCMs) provided by participating regional and international research institutions. Additional information on the current state of the climate system, and forecasters' previous experiences with similar meteorological situations are also incorporated into the final product. The final probabilities for each category given to a region are agreed by a consensus process among the participating experts. The contingency table approach is the most popular method for obtaining first estimates of probabilities because of its intuitive appeal. However, this method is known to estimate probabilities unreliably (Mason and Mimmack 2002), and so there is no guarantee that the consensus-building revisions to the forecasts will show high levels of reliability either.

b. Defining the target variable

RCOF forecasts typically are presented as maps showing probabilities of seasonal accumulations (in the case of precipitation), or averages (in the case of temperature) falling within predefined categories. However, it is not always clear whether these accumulations or averages relate to areal averages, and if they do, it is not always clear what the area is over which the target variable is to be averaged. For example, consider the idealized example shown in Figure 1, in which there are forecasts of seasonal rainfall totals for three regions. The forecasts for regions I and II were constructed by calculating a regional average rainfall index, and then forecasting the index. For region III, however, the forecast was constructed by calculating two separate regional indices, whose areas are delimited by the dashed line, and then combining the two regions because the forecasts were identical, or at least very similar. The problem now, however, is that the three forecasts no longer mean the same thing: the forecasts for regions I and II define probabilities for seasonal rainfall totals averaged over the respective regions, but the forecast for region III does not mean that the seasonal rainfall total averaged over this region has a 25% chance of being above-normal. Instead, for region III, the probability of the seasonal rainfall total averaged over sub-region *a* has a 25% chance of being above-normal, and the same is true of sub-region *b*.

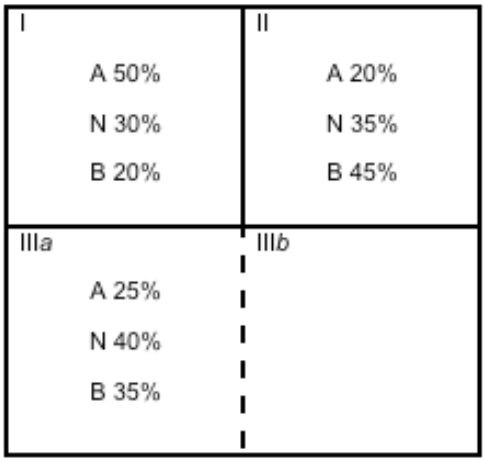


Figure 1: Idealized example of seasonal rainfall forecasts for three regions. A indicates the probability of above-normal rainfall, N of normal, and B of below-normal.

Why is this difference in interpretation important? Imagine a situation in which observed rainfall over sub-region *a* is above-normal and over region *b* is below-normal, and that the spatial average over the whole of region III is then normal; this forecast would be scored based on the 40% for the verifying category (normal). If the sub-regions had been left as separate forecasts it would be scored based on the 25% for the above-normal category over sub-region *a*, and the 35% for the below-normal category over sub-region *b*, and so would score worse, and appropriately so. The point is that the distribution of observed rainfall over region III (wet on the one half, and dry on the other) should result in a forecast that scores poorly, but only because the forecasts for the two sub-regions were similar does the forecast score much better.

This situation of the forecast verification being sensitive to the degree to which regions are combined because of similar predictions has to be considered undesirable. If the forecasts for sub-regions *a* and *b* had been slightly different such that their areas had not been combined, surely it would then be unfair that the forecast map verifies so much more poorly. It seems to make sense to verify the sub-regions separately even if the forecasts are the same. Unfortunately, in most cases, these sub-regions are not indicated on the map, and so it may be impossible to identify where they are, and how many there are. In drawing up the consensus, the boundaries for the original sub-regions may have been modified anyway.

A more serious problem is that it is no longer possible to identify exactly what the forecast for any region means: given a forecast like that shown in Figure 1, but without the dotted line, how is one to know that the forecast for region III does not mean that the rainfall averaged over this entire region has a 25% probability of being above-normal? And without any information about the sub-regions, what can one conclude that the forecast means anyway? This problem of the interpretation of forecasts for regions is not only a problem for verification analyses, it is also a problem for the users of the forecasts, and needs to be addressed as an issue in the construction of the forecast. Possible solutions include providing forecasts on a gridded basis [as followed by Global Producing Centres (GPCs), for example], or indicating the sub-regions on the map as thin lines, or forecasting for stations rather than for regional indices. Forecasting for pre-defined homogeneous zones is also an attractive option. A solution to this problem needs to be identified, and implemented at the RCOFs.

A further problem arises when verifying forecasts of regional indices if the regions are not the same size. Supposing that the problem of the sub-regions can be resolved, if the regions are of differing size, the verification results should reflect this fact. For example, imagine a simple example where there are only two regions, one three times as large as the other. If two forecasts issued a 50% probability on the verifying category for one of the regions, and a 20% probability on the other, the forecast that had the 50% over the larger region should surely be scored better than the other. The verification results should be weighted by area to resolve this problem.

One way to resolve the problems of the sub-regions and of unequally sized regions is to grid the forecasts. This solution is particularly attractive if the verification data themselves are gridded, and has been adopted in the analyses presented here.

3. Verification of RCOF Forecasts

a. Why Verify Forecasts?

The RCOFs have been very successful in developing the capacity of meteorological services throughout Africa and beyond to produce operational seasonal forecasts, but the application of this new information still remains highly limited partly because of a lack of information about the quality of the forecasts. Understanding the quality of the forecast allows for correct interpretation and effective use of the information. The objectives for evaluating the quality of the forecasts were categorized by Brier and Allen (1951) as serving administrative, scientific and economic purposes. The main objectives for this evaluation are administrative and scientific. One aim is to monitor the performance in order to inform the current RCOF process of any bias in the forecast system, which should subsequently lead to improvements in the forecast methodology. A second aim is to ensure that users of seasonal forecasts have appropriate levels of confidence in the forecasts they use for them to be able to optimally utilize the information. Users are ultimately interested in the value of the forecast information, which is distinct from its quality: value pertains to the incremental benefits derived from the use of a seasonal forecast whether these benefits are economic, social or otherwise, whereas quality pertains to the correspondence between what was forecast and what was observed. Although an assessment of the quality of the forecasts does not address the question of whether they are actually useful, seasonal forecasts can have no value if their quality is bad, i.e. if they provide no indication of how the climate of the coming season may differ from that of previous years. Therefore evaluating the quality of the forecast becomes of paramount importance before considerations of assessing the value of the forecast.

b. Methods for Verification

Measuring the quality of probabilistic forecasts is much more complicated than for deterministic forecasts. Consider the simple example of forecaster A, who says it is going to rain tomorrow, and forecaster B, who says there is a 60% chance of rain tomorrow. If it rains, forecaster A clearly issued a correct forecast, but what about forecaster B? And is forecaster B correct or incorrect if it does not rain? To forecaster B, it does not seem to matter whether it rains or not, (s)he has not made an incorrect forecast. The temptation is to conclude that probabilistic forecasts cannot be “wrong” (as long as probabilities of 0% are never issued to any of the possible outcomes), and that therefore these forecasts are always correct. While this conclusion is logically valid, it is, of course, also distinctly unhelpful since any probabilistic forecast that does not issue zero probabilities is as equally “correct” as any other. Probabilistic forecasts can only be described as “correct” in the sense that they indicated that the observed outcomes could have happened, in which case the probabilities themselves become completely irrelevant (again, as long as they are greater than zero). The question of correctness of probabilistic forecasts is then so uninformative as to be essentially useless, and nothing is learned about whether the forecasts have successfully indicated whether or not the observed outcomes were likely or unlikely to have happened. Therefore more meaningful questions about the quality of probabilistic forecasts need to be asked.

One reasonably common practice is to define probabilistic forecasts as “correct” if the category with the highest probability verified. Hit scores are then calculated, and have a reasonably intuitive interpretation as long as the user has a good understanding of the base

rate.[1] Sometimes "half-hits" are counted if two of the categories (in a three-category system) have increased probabilities above climatology and one of these two categories verifies. There are some inter-related and important problems with such approaches. Firstly, the verification procedure implicitly condones the interpretation of the forecast in a deterministic manner, which is a problem both for the user (who loses information about the uncertainty in the forecast), and for the forecaster (who typically becomes tempted to hedge towards issuing higher probabilities on the normal category to avoid a two-category "error"). Secondly, if the probabilities are to be considered as at all meaningful, one does not actually want to achieve a high score because that would indicate that the forecasts are unreliable. If the highest probability is 40% one should want a "hit" only 40% (i.e. less than half) of the time. Thirdly, the scoring system does not give any credit for issuing sharp probabilities. Thus two forecasters who always issue the same tendencies in their forecasts will score exactly the same regardless of whether one of the forecasters is more confident than the other.

Rather than trying to transform the forecasts so that individual forecasts can be counted as "correct" or "incorrect" in some way, it is recommended that verification procedures be used that are suitable for the forecasts in the format in which they are presented. A draft set of scores and procedures suitable for the verification of RCOF-type forecasts has been submitted to the WMO CCl for review (Mason 2009). A subset of these procedures is applied in this analysis.

Reliability diagrams

For detailed information on forecast quality, the reliability diagram is recommended (Mason 2009). The diagrams provide useful indications of most of the important attributes of forecast quality. Reliability diagrams are based on a diagnosis of probabilistic forecasts for a predefined set of events, and so can be constructed for each of the categories separately. However, it is not required that the definition of an event remain fixed, and so a single diagram can be constructed for all the categories combined. Both approaches are recommended, since the diagrams for the individual categories are useful for indicating whether the quality of the forecasts depends on the outcome, while the combined diagram is useful for examining whether the probabilities can be interpreted consistently across the categories.

The basic idea of the reliability diagram is very simple, but the diagram contains a wealth of information about the quality of the forecasts. For each discrete value of the forecast probability, the reliability diagram indicates whether the forecast event occurred as frequently as implied. The different forecast probabilities are plotted on the *x*-axis, and on the *y*-axis the "observed relative frequency" of the event is shown. The observed relative frequency for the k^{th} forecast probability, $\overline{y}_k$, is the number of times an event occurred divided by the number of times the respective probability value was forecast:

$$\overline{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{k,i} \ . \qquad (1)$$

where n_k is the number of forecasts of the k^{th} forecast probability, and $y_{j,i}$ is 1 if the i^{th} observation was an event, and is 0 otherwise. Equation (1) is akin to a hit score, which is the number of hits divided by the number of forecasts.

[1] Knowledge of the base rate is necessary because the naïve expectation is that at least 50% of the forecasts should be correct. However, with three or more categories scores of less than 50% correct may be good.

It is common practice to include a histogram showing the frequency of forecasts for each point on the curve. The histograms are useful for indicating any unconditional biases in the forecasts, and they also show the forecast sharpness. Histograms for forecasts that have no unconditional bias will be centred on the relative frequency of the event over the verification period. Forecasts with weak sharpness have histograms that have high frequencies on probabilities (typically close to the climatological probability. Sharp forecasts have histograms showing high frequencies of forecasts near 0 and 100% – the histograms are *u*-shaped. For seasonal forecasts, *u*-shaped histograms are exceptionally rare because of an inability to be so confident, but relatively sharp forecasts have more dispersed histograms.

The reliability curve itself can be deceptively difficult to interpret because it does not represent the frequency of forecasts on each probability value. Sampling errors can therefore vary quite markedly along the curve. It is recommended that least squares regression fits to the curves be calculated, weighted by the frequency of forecasts on each probability, and added to the diagrams (Wilks and Murphy 1998). The parameters of the regression fit can be estimated using

$$\beta_1 = \frac{\sum_{k=1}^{d} n_k \left(p_k - \overline{p}\right)\left(\overline{y}_k - \overline{y}\right)}{\sum_{k=1}^{d} n_k \left(p_k - \overline{p}\right)^2}. \qquad (2a)$$

and

$$\beta_0 = \overline{y} - \beta_1 \overline{x}. \qquad (2b)$$

where β_1 is the slope and β_0 the intercept of the fitted regression line, d is the number of discrete probability values, n_k is the number of forecasts for the k^{th} probability value, $\overline{p}_k$ is the k^{th} probability value, $\overline{p}$ is the average probability, $\overline{y}_k$ is the observed relative frequency for the k^{th} probability value [Eq. (1)], and $\overline{y}$ is the observed relative frequency over the verification period. It is recommended that the slope of the regression line, which can be viewed as a measure of resolution, be communicated as a percentage change in the observed relative frequency given a 10% increase in the forecast probability. If the forecasts have good resolution an event should increase in frequency by 10% as the forecast probability is incremented by each 10% (e.g., from 30% to 40%, or from 40% to 50%), and the slope will be 1.0, but if they have no resolution the slope will be zero. Over-confidence will be indicated by a slope of between 0.0 and 1.0 (the increase in frequency will be between 0% and 10%), while under-confidence will be indicated by slopes of greater than 1.0 (increases in frequency of more than 10.

One major limitation of reliability diagrams is that they require a large number of forecasts because of the need to calculate the observed relative frequencies for each forecast value. The diagrams can therefore only be constructed by pooling forecasts for different years and locations. Reliability diagrams are therefore drawn only for each region and target season rather than for individual locations.

ROC Graphs

The ROC graph is constructed by calculating the ability of the forecasts to successfully identify the events. Starting with the forecasts with highest probabilities, the observations that are most confidently indicated as events are highlighted. The events that are selected are called “hits”. The proportion of all events thus selected is calculated, and is known as the hit rate (HR), or probability of detection (POD):

$$\mathrm{HR} = \frac{\text{number of hits}}{\text{number of events}}. \tag{3}$$

It is possible that some non-events have been selected incorrectly, and these are known as "false-alarms". The proportion of non-events incorrectly selected [the false-alarm rate (FAR)] is calculated also:

$$\mathrm{FAR} = \frac{\text{number of false-alarms}}{\text{number of non-events}}. \tag{4}$$

The hit and false-alarm rates are commonly tabled, and given the general practice in seasonal forecasts that probabilities are rounded to the nearest 5% (except for the climatological probability), it is recommended that the table be constructed for each discrete probability value.

If the forecasts have no useful information, the hit and false-alarm rates will be identical, but if the forecasts can discriminate the events, the hit rate will be larger than the false-alarm rate. Since it is unlikely that all the events were correctly selected using only the forecasts with highest probabilities, additional selections are made using the next highest probability, and the hit and false-alarm rates are updated. The difference between the hit and the false-alarm rate is expected to be a little less than at the first step, since we are less confident about having correctly selected the events. These steps are continued until all the events have been selected. The hit rates are then be plotted against the false-alarm rates.

To calculate the area beneath the curve by the trapezoidal rule, the following equation can be used:

$$A = 0.5 \times \left[1 + \sum_{k=0}^{d} \left(y_k x_{k+1} - y_{k+1} x_k \right) \right]. \tag{5}$$

where d is the number of discrete probability values, and y_1 and x_1 are the hit and false-alarm rates for the highest probability value only, y_2 and x_2 are the rates for the highest and second highest probabilities, etc.. For $i = 0$ the hit rate and false-alarm rates are defined as 0.0, and for $i = m + 1$ they are defined as 1.0 to ensure that the curve starts in the bottom-left, and ends in the top-right corners, respectively.

Generalized discrimination

The generalized discrimination score (Mason and Weigel 2009) provides an indication of the ability of the forecasts to discriminate wetter observations from drier. It is a multi-category version of the area beneath the ROC graph. The generalized discrimination score, D, is defined as

$$D = \frac{\sum_{k=1}^{m-1} \sum_{l=k+1}^{m} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} I\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right)}{\sum_{k=1}^{m-1} \sum_{l=k+1}^{m} n_k n_l}, \tag{6a}$$

where m is the number of categories, n_k is the number of times the observation was in category k, $\mathbf{p}_{k,i}$ is the vector of forecast probabilities for the i^{th} observation in category k, and

$$I\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) = \begin{cases} 0.0 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) < 0.5 \\ 0.5 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) = 0.5 \\ 1.0 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) > 0.5 \end{cases}, \tag{6b}$$

and where

$$F\left(\mathbf{p}_{k,i},\mathbf{p}_{l,j}\right)=\frac{\sum_{r=1}^{m-1}\sum_{s=r+1}^{m}p_{k,i}(r)p_{l,j}(s)}{1-\sum_{r=1}^{m}p_{k,i}(r)p_{l,j}(r)}, \qquad (6c)$$

where $p_{k,i}(r)$ is the forecast probability for the r^{th} category, and for the i^{th} observation in category k.

Although Eq. (6) may seem complicated, its interpretation is fairly simple: What is the probability of successfully discriminating the wetter (or warmer) or two observations? Equation (6a) compares each of the observations in the normal and above normal categories with each of those in the below-normal category in turn, and calculates the probability that the forecasts correctly point to the observation in the normal or above-normal category as the wetter (warmer). This procedure is then repeated comparing each of the observations in the normal category with each of those in the above-normal category. The selection of the wettest observation is based on Eq. (6c), which defines the probability that a value randomly drawn from $\mathbf{p}_{l,j}$ will exceed one randomly drawn from $\mathbf{p}_{k,i}$. If this probability is greater than 0.5 [Eq. (6b)] the forecasts suggest that it is more likely that the second observation (that corresponding to $\mathbf{p}_{l,j}$) is wetter (or warmer) than the first (that corresponding to $\mathbf{p}_{k,i}$).

The generalized discrimination score can be calculated for each location, and then a map of the scores can be drawn to indicate areas in which the forecasts have some resolution. The score has an intuitive scaling that is appealing to many non-specialists: it has an expected value of 50% for useless forecast strategies (guessing, or always forecasting the same probabilities), and good forecasts will have a score greater than 50%, reaching 100% given perfect discrimination. Scores of less than 50% indicate bad forecasts (forecasts that can discriminate, but which indicate the wrong tendency – for example, high forecast probabilities on below-normal indicate a low probability that below-normal rainfall will actually occur), and can reach a lower limit of 0% given perfectly bad forecasts. The score represents the most meaningful answer to the naïve question: “How often are the forecasts correct?” without having to reduce the forecasts to a deterministic format. One major problem with the score is that it is insensitive to the reliability of the forecasts, and so is unaffected by any monotonic transformation of the forecasts. This problem is considered less severe than for the insensitivity of the area beneath the ROC graph (discussed below) to reliability of the forecasts because of the way in which the probabilities are compared across the categories in the generalized discrimination score. With the ROC area, for example, all the probabilities could be multiplied by a factor k, where $0 < k < 1$, and the area is unaffected, but since the probabilities across the categories have to add to 1.0, simple rescalings cannot be applied without affecting the probabilities in the other categories, and thus also affecting the probability of successfully discriminating observations.

For a number of reasons the generalized discrimination score is recommended in place of the more commonly used ranked probability skill score (RPSS) as a summary measure of skill to indicate where the forecasts are good. Firstly, the RPSS does not have any intuitive interpretation, which essentially renders it an abstract number to most non-specialists. Secondly, its scaling can be confusing: while positive values indicating skill should be simple enough to understand, the fact that it does not have a lower bound of -100% means that the score is asymmetric, so that forecasts with a score of -50% are not as bad as forecasts with a score of 50% are good. But even the idea of having 0% as no skill rather than 50% seems much more logical to verification experts than it does to users with only weak mathematical backgrounds.

The main reason for not recommending the RPSS is that it is frequently misinterpreted even by forecasters, and typically results in a more pessimistic view of the quality of the forecasts than is warranted. There is a widespread belief that if the score is less than zero the forecasts are worse than climatological forecasts, and the user would therefore have been better off with the latter, but this is not necessarily the case. Consider a set of 10 forecasts for a single location, five of which indicate a probability of 60% for above-normal rainfall, and the other five only a 10% probability. The climatological probability of above-normal rainfall is 33%. For the sake of simplicity, the below-normal and normal categories can be combined. Now imagine that for two of the years for which a 60% probability was issued rainfall was above-normal; 40% of these years were thus above-normal, and the forecasts have successfully, but over-confidently, indicated an increase in the probability of above-normal rainfall. For the years in which a 10% probability was issued, only one of these was above-normal; thus 20% of these years were above-normal, and the forecasts have successfully, but over-confidently, indicated a decreases in the probability of above-normal rainfall. These forecasts appear to contain useful information because they have correctly indicated increases and decreases in the probability of above-normal rainfall. However, the RPSS (which in this case is equivalent to the Brier skill score, because the normal and below-normal categories are combined) is about -7%, indicating negative skill. The score is slightly worse still if it is objected that above-normal rainfall was observed only 30% of the time over the 10-year period rather than 33%. The problem is that the RPSS has an arbitrary requirement that the resolution of the forecasts must be greater than the errors in the reliability, and since these forecasts show marked over-confidence they score badly[2]. By trying to measure resolution and reliability at the same time, the score becomes difficult to interpret, whereas the generalized discrimination score measures only the one attribute, and thus provides a simpler indication of whether the forecasts might be useful.

c. Data

Monthly rainfall raingauge- and satellite-based observations from the NOAA NCEP CPC CAMS OPI (Janowiak and Xie 1999) and UEA CRU TS 2.1 (Mitchell and Jones 2005) datasets were used. The satellite data were chosen because of their spatial and temporal coverage as opposed to actual gauge precipitation data, which are unevenly distributed and intermittently recorded across the regions. Two precipitation datasets were used because the one does not extend back far enough to calculate the climatology of the regions, and the other does not extend ahead enough to cover the forecast verification periods. The UEA data were spatially aggregated to the coarser resolution of the CAMS OPI data (2.5°×2.5°), and then were transformed using the following procedure to eliminate inconsistencies between the two datasets.

1) Calculate the parameters of a gamma distribution fitted to the UEA data (UEA) for the period of overlap of the two datasets (1979 – 2002).
2) For the UEA data prior to the period of overlap (1961 – 1978), transform the data to quantiles of the corresponding gamma distribution using the parameters obtained from step 1.

[2] Some attempts have been made to correct these so-called biases in the ranked probability skill score by introducing an adjustment for the uncertainty in the climatological probabilities, but the corrections are applicable only for ensemble prediction systems, and so it is not clear how they could be applied for consensus forecasts. These debiased scores are considered useful in the context of the CBS SVSLRF, which targets GPC products, but cannot be applied in the current context. Besides, the criticism remains that these scores are still abstract numbers, and so are difficult to understand by all but specialists in forecast verification.

3) Calculate the parameters of a gamma distribution fitted to the CAMS OPI data for the period of overlap (1979 – 2002).
4) Transform the quantiles from step 2 to deviates using the gamma distribution parameters for the CAMS OPI data obtained from step 3.
5) Append the CAMS OPI data onto these transformed data. The first part of the climatological period now consists of transformed UEA data, and the second part consists of data from the CAMS OPI dataset that has not been transformed.

This transformation procedure should eliminate most biases resulting from differences in means and variances of the two datasets.

The duration of the seasonal forecast information and the spatial coverage of the rainfall data used for verification are as follows:

RCOF	Verification Periods	Lead-Times (months)	Area Covered
SARCOF	OND 1997 – 2007 JFM 1998 – 2007	0-1 0 and 3-4	7.5°N – 35.0°S, 10.0°E – 60.0°E
GHACOF	MAM 1998 – 2007 SOND 1998 – 2007	0 0	25.0°N – 12.5°S, 20.0°E – 52.5°E
PRESAO	JAS 1998 – 2007	1	25.0°N – 0.0°, 17.5°W – 25.0°E

In any verification exercise the quality and quantity of the sample data determines the quality and robustness of the verification results. Reliable results are usually associated with large samples both spatially and temporally. It has to be acknowledged here that the approximately 10 years of data provide a rather small sample size. The gridding of forecasts and pooling of the grid points implemented in this exercise is therefore meant to improve the sample size and provide a reasonable indication of RCOF forecast performance.

The forecast maps were gridded to the same resolution (2.5°×2.5°) as the estimated rainfall to allow for a grid-point comparison of the forecasts and the corresponding observed rainfall. A total number of 142 grid points were used in this analysis for Southern Africa; 76 for the Greater Horn of Africa and 86 for West Africa. The number of grid points varied from one year to the other depending on the aerial extent of the forecast issued in a particular year.

4. Results

a. Southern Africa

i. October – December

Attributes diagrams for forecasts of above-normal rainfall over southern Africa for the season October to December are shown in Figure 1. In general, reliability for the above-normal and below-normal categories is moderate, with the category increasing in frequency by about 5%, and 6%, respectively, for every 10% increase in the forecast probability. However, the reliability curves are clearly not very straight, which is partly a result of a small sample size, but which also indicates that the forecasts are not well-calibrated. For the above-normal category, for example, the reliability curve is essentially flat (indicating no resolution) except for forecast probabilities of 45% and higher. The ROC curves (black and light grey lines on Figure 2) confirm that the skill for these two categories is weak, and that skill comes primarily from the forecasts for above-normal rainfall with probabilities of 45% and higher (the curve is steeper than 45° in the lower left corner, but then follows the

diagonal to the top right). However, the successful discrimination rates for the below- and above-normal categories are similar (about 54% and 55% respectively). For the normal category, there is no skill at all (dashed curve on Figure 2; successful discrimination rate of 50%), and the reliability curve is essentially flat.

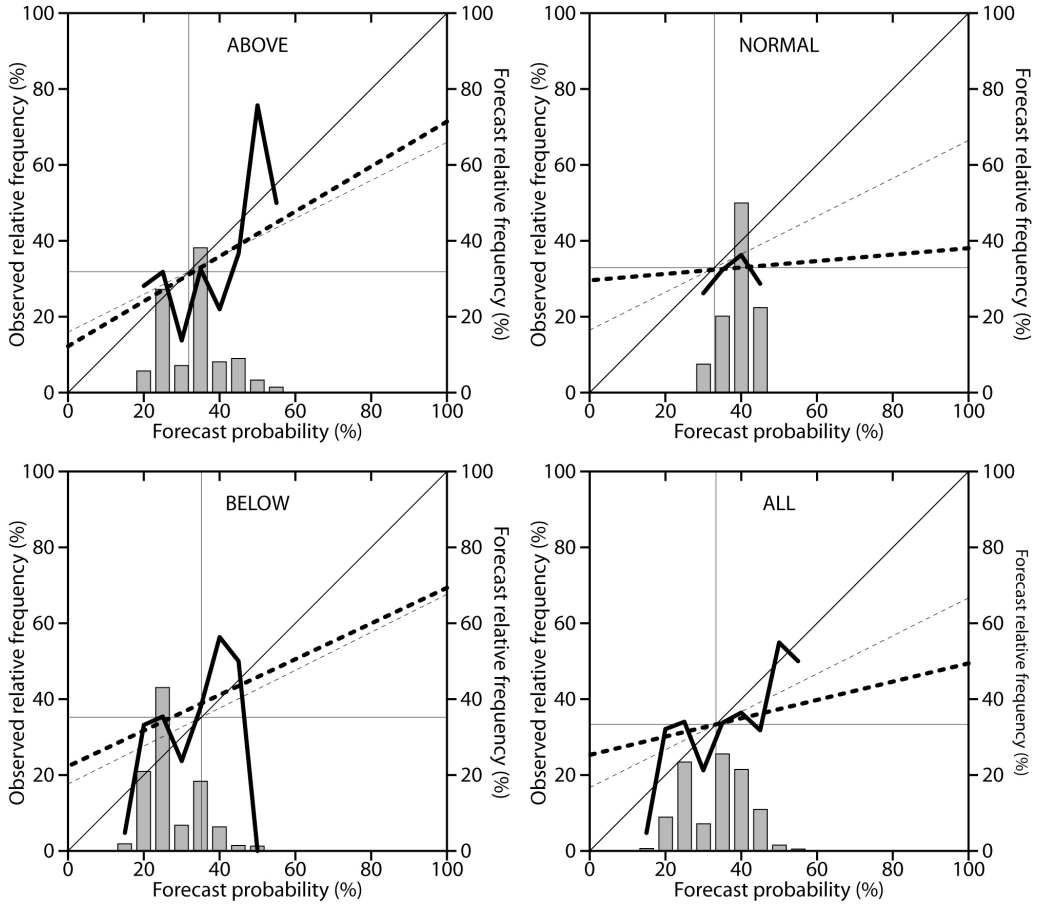


Figure 1. SARCOF attributes diagrams for the October to December season. The thick black line shows the reliability curve, and the thick dashed line is the least squares weighted regression fit to the reliability curve. The weights are shown by the grey bars, which indicate the relative frequency of forecasts in each 5% bin. The thin horizontal and vertical lines indicate the relative frequency of occurrence of rainfall in the respective category, while the thin diagonal represents the line of perfect reliability, and the thin dashed line the line of "no skill" as measured by the Brier score.

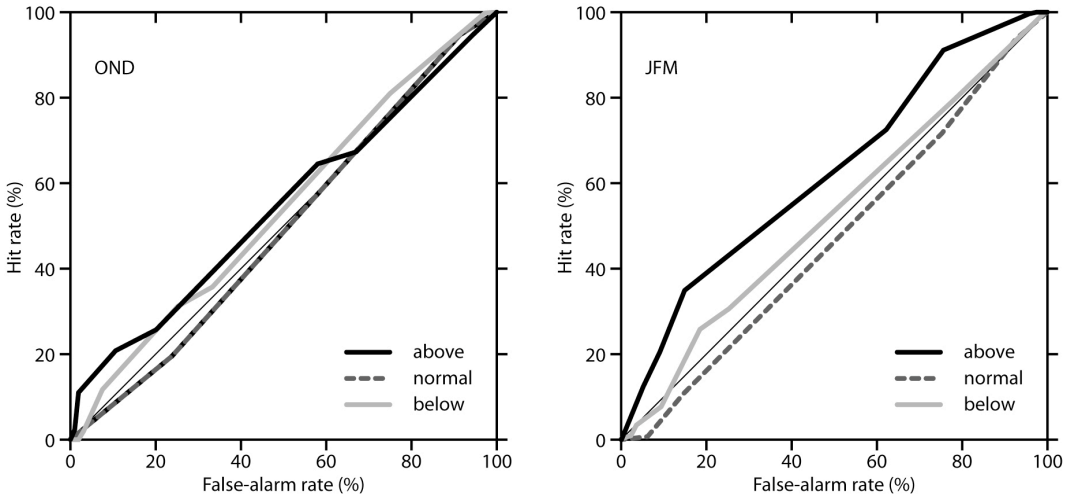


Figure 2. ROC diagrams for the October to December (OND) and January to March (JFM) SARCOF forecasts. The thick black line is for above-normal rainfall, the dashed grey line for normal, and the light grey line for below-normal.

Despite the poor reliability, the biases in the forecasts are relatively minor compared to those for the RCOFs in West Africa and the Greater Horn: below-normal rainfall is under-forecast (the average forecast probability is 27% compared to an observed relative frequency of 36%). This under-forecasting was primarily compensated by over-forecasting of the normal category (the average forecast probability was 39%, but this category was not observed more frequently than expected climatologically). The high frequency of forecasts of the normal category with probabilities above climatology is suggestive of hedging.

Overall, given the general lack of resolution for the less sharp forecasts, including the complete lack of resolution for the normal category, on average a 10% increase in forecast probability translates to only an approximately 2% increase in the probability of occurrence. A positive outcome is that there is little bias in the forecasts, but the verification period was not noticeably different from the climatological period. As a result, the October – December forecasts for southern Africa show only weak positive skill (although the generalized discrimination suggests marginally negative skill at 48%), most of which comes from the sharpest probabilities (15% and less, and 50% and more). There is no obvious spatial distribution of skill (Figure 3), although there are weak indications of greater skill in south-eastern parts of the mainland compared to elsewhere.

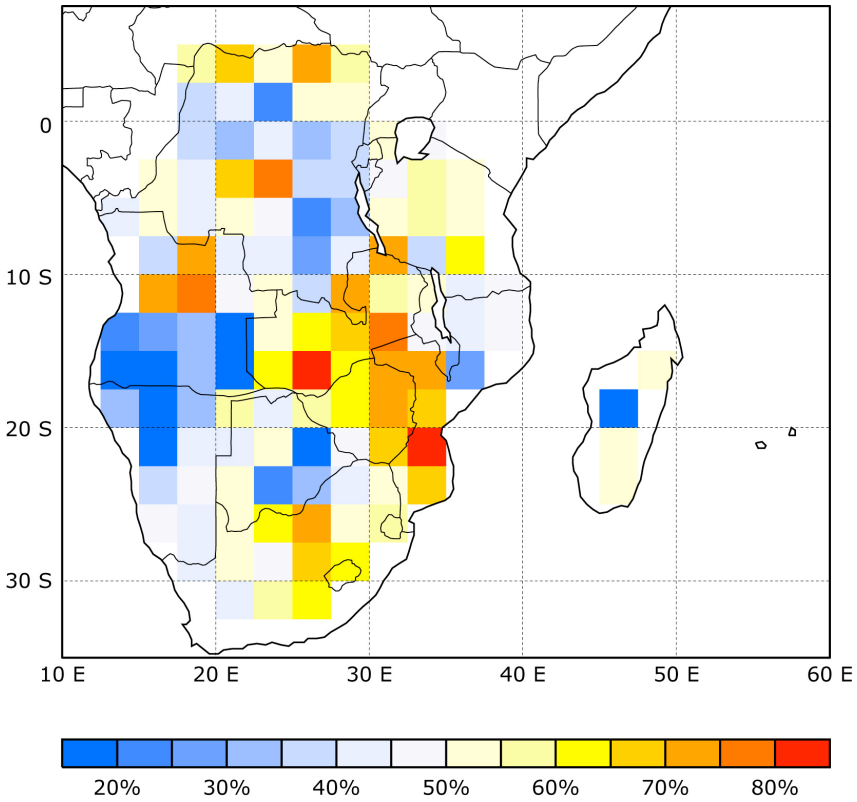


Figure 3. Map of generalized discrimination for the SARCOF October – December forecasts.

ii. January – March

Forecasts of above-normal rainfall over southern Africa for the season January to March show better reliability and resolution than for October – December (Figure 4). Although the reliability curve is again somewhat noisy, the ROC graph indicates much better discrimination (62%; Figure 2) than for any of the forecasts for October – December. In general the occurrence of above-normal rainfall increases by about 11% for every 10% increase in the forecast probability, suggesting slight under-confidence, but there is clear evidence of over-forecasting: over the verification period the average probability for above-normal was about 34% whereas above-normal occurred only about 26% of the time. The over-forecasting is most clearly evident for the lowest forecast probabilities, which contributes to the slight under-confidence: generally, when notable decreases in the probabilities of above-normal rainfall are indicated (probabilities of 25% or less), these decreases are underestimated.

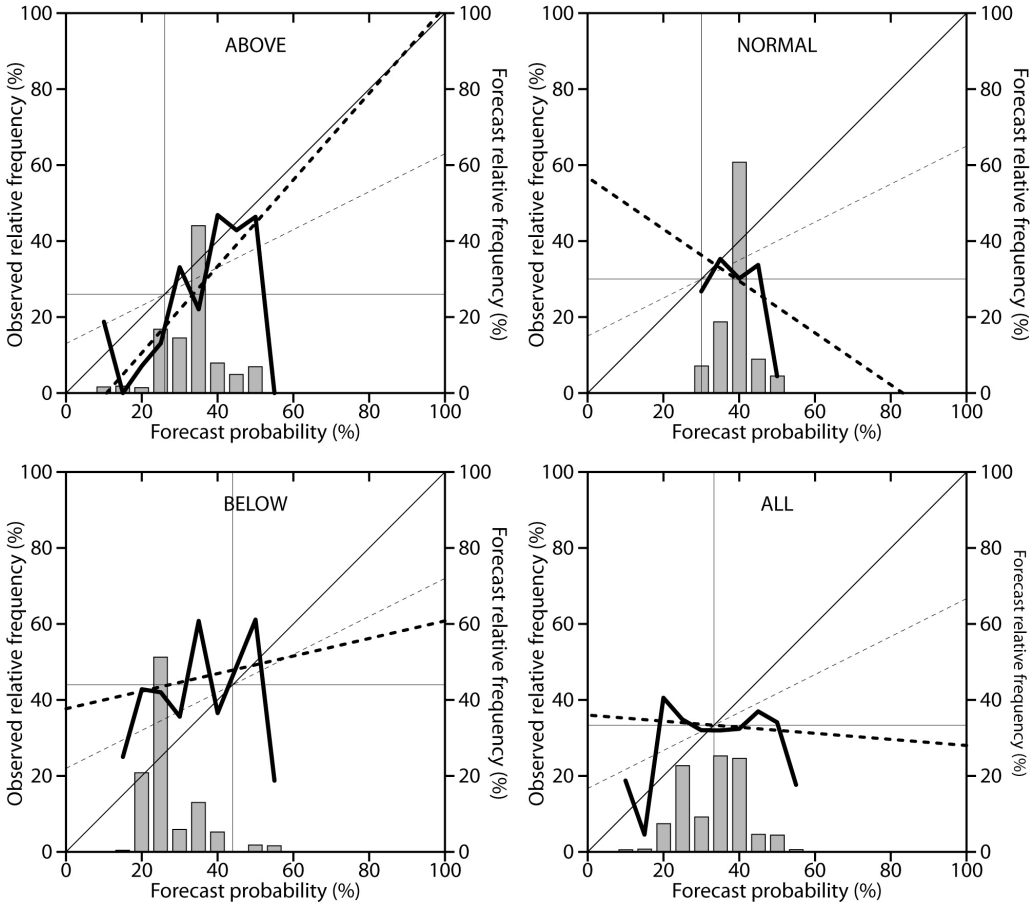


Figure 4. SARCOF attributes diagrams for the January to March season. The thick black line shows the reliability curve, and the thick dashed line is the least squares weighted regression fit to the reliability curve. The weights are shown by the grey bars, which indicate the relative frequency of forecasts in each 5% bin. The thin horizontal and vertical lines indicate the relative frequency of occurrence of rainfall in the respective category, while the thin diagonal represents the line of perfect reliability, and the thin dashed line the line of "no skill" as measured by the Brier score.

There is much less skill for the below-normal category compared to the above-normal (the forecasts discriminate below-normal rainfall with only a 52% success rate; Figure 2): these forecasts are notably over-confident (below-normal rainfall increases only by about 2% for every 10% increase in the forecast probability), and there is a marked bias with below-normal conditions being under-forecast (the average forecast probability was about 27% compared to an observed frequency of almost 44%). The over-forecasting of the above-

normal category and under-forecasting of the below-normal represents a failure to indicate a fairly marked shift in the climatology of the region towards drier conditions during the verification period compared to 1961 – 1990. There have been no forecasts for below-normal rainfall with probabilities exceeding 40% since the El Niño of 1997/98, but probabilities of 25% have far outnumbered all other forecasts, and yet the forecasts have no resolution within this range. It is possible that there is a greater reluctance to forecast high probabilities of below-normal than of above-normal since in many parts of the region a warning of dry conditions would be considered more serious than a warning of wet. Forecasts of 25%, 40%, 35% (for below-normal, normal, above-normal, or other probabilities with the same ranking of the categories) were issued about four times as frequently as probabilities of 35%, 40%, 25% (or similar ranking).

As for the October – December season, the over-forecasting of the normal category (average forecast probability was 39% compared to a 30% relative frequency of occurrence) is suggestive of hedging. The normal category also has markedly poor resolution (normal rainfall decreases in frequency by about 7% for every 10% increase in the forecast probability), largely because of the low observed relative frequencies for forecasts with the highest probabilities (50%). Because these forecasts were issued relatively infrequently, the skill is only weakly negative (47% correct discrimination; Figure 2). The forecasters frequently issue probabilities of 40% for the normal category, but should be made aware that there is no evidence of an ability to forecast an increase in the probability of this category above its climatological value. The lack of skill for the normal category is not exclusive to the SARCOF forecasts, or to those of the other African RCOFs, but is widely reported elsewhere (Wilks 2000; Wilks and Godfrey 2002).

There is stronger indication of a spatial distribution to the skill of the forecasts than for October – December, with better skill south of about 10°S (Figure 10). Predictability in the vicinity of Malawi is known to be weak because of a transition between zones with distinct ENSO-teleconnection signals to the north-east (generally wet during warm episodes), and south-west (generally dry).

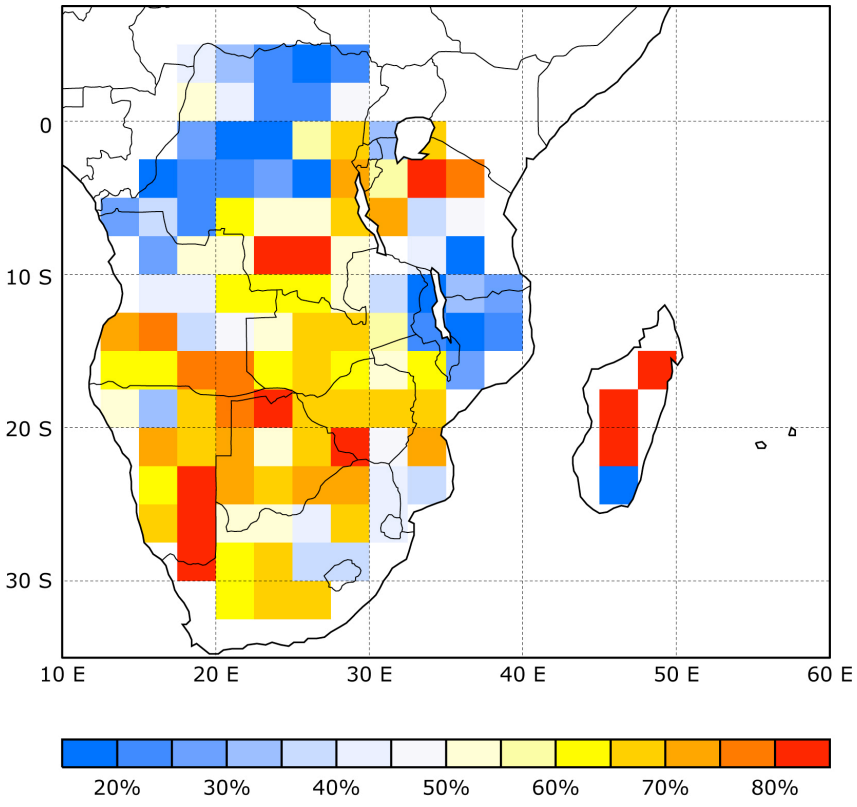


Figure 5. Map of generalized discrimination for the SARCOF January – March forecasts.

Overall, the forecasts for January – March are notably more skilful than those for October – December, although the generalized discrimination is only marginally better (49% compared to 48%) because of the strong biases. The skill for January – March is largely contributed by good forecasts of the above-normal category; for the below-normal category the forecasts have poor reliability and resolution, and for the normal category there is no skill at all. Hedging on the normal category contributed to the under-forecasting of the below-normal category (below-normal rainfall occurred more widely than suggested by the forecasts), but the general tendency of the forecasts over the ten-year verification period was incorrect anyway.

b. Greater Horn of Africa

i. March to May

The forecasts for the March to May period for the Greater Horn of Africa show poor discrimination (Figure 6), with correct rates of only 56%, 49%, and 53% for the above-normal, normal, and below-normal categories, respectively. Most of the skill that is evident appears to come from forecasts for Kenya, and neighbouring areas (Figure 7). Accordingly, the forecasts show poor resolution and reliability (Figure 8), although this season in this region is generally considered one of relatively poor predictability. There are increases in frequency of the above-, normal, and below-normal categories by only about 3%, 0%, and 4% for every 10% increase in the respective forecast probabilities.

In addition to the poor resolution, there are some serious biases in the forecasts, which failed to indicate a predominance of below-normal rainfall over the verification period (below-normal rain was observed 55% of the time, but the average forecast probability was only 30% for this category). This under-forecasting was at the expense of notable over-forecasting of both the normal and the above-normal categories: the average forecast probability for normal was 40%, but normal rainfall occurred only 31% of the time, while the corresponding values for the above-normal category were 30% and 14%. Thus above-normal rainfall occurred less than half as often as indicated by the forecasts. The infrequent occurrence of above-normal rainfall over the verification period represents a marked difference from the climatological period, and a failure of the forecasts to provide any indication of this shift has to be considered a major weakness.

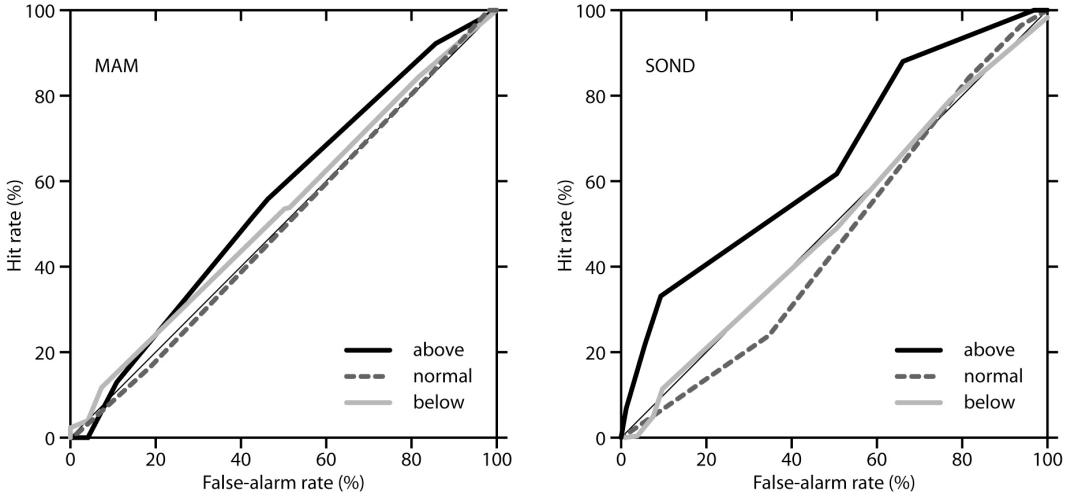


Figure 6. ROC diagrams for the March to May (MAM) and September to December (SOND) GHACOF forecasts. The thick black line is for above-normal rainfall, the dashed grey line for normal, and the light grey line for below-normal.

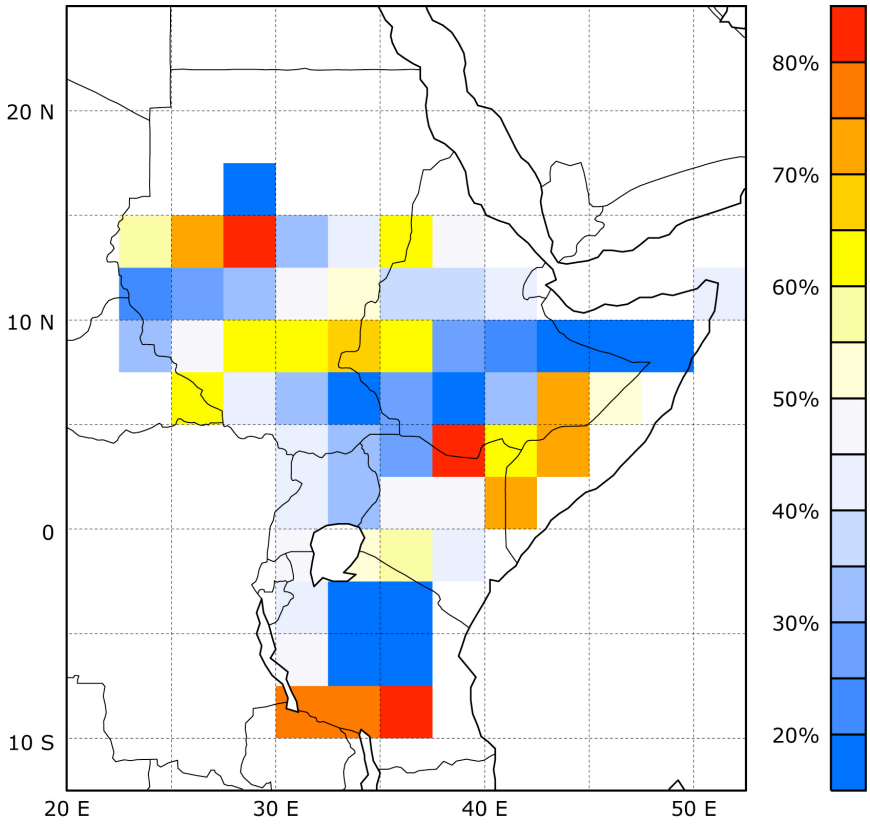


Figure 7. Map of generalized discrimination for the GHACOF March – May forecasts.

As with the SARCOF forecasts, the evidence for hedging is clear. Two-thirds of all the forecasts for this season indicated a 40% chance of normal rainfall, and for more than half of the remaining times the probability was even higher. These probabilities are consistently too high, and the complete lack of skill for this category does not justify increasing the probability above its climatological value anyway. If it were not for the few cases of 30% forecast probability on normal (the lowest probability ever assigned to this category) when normal rainfall did not occur at all, the observed relative frequency of normal rainfall would have decreased monotonically with increasing probability.

Overall the March – May forecasts for the Greater Horn of Africa, show no skill and poor resolution (the generalized discrimination is 44%, and a 10% increase in forecast probability implies only a 1% increase in observed relative frequency). Both the generalized discrimination and the overall resolution are affected by the biases; if these biases are ignored, the forecasts for the below-normal and above-normal categories do have very weak positive skill, thus increases or decreases in probability within a category are meaningful, but the probabilities between the categories cannot be compared, which would make the forecasts very difficult to use profitably. The biases are most severe for the outer categories, with below-normal (above-normal) rainfall occurring far more (less) frequently and extensively than forecast. Hedging on the normal category does not help the bias on the below-normal category.

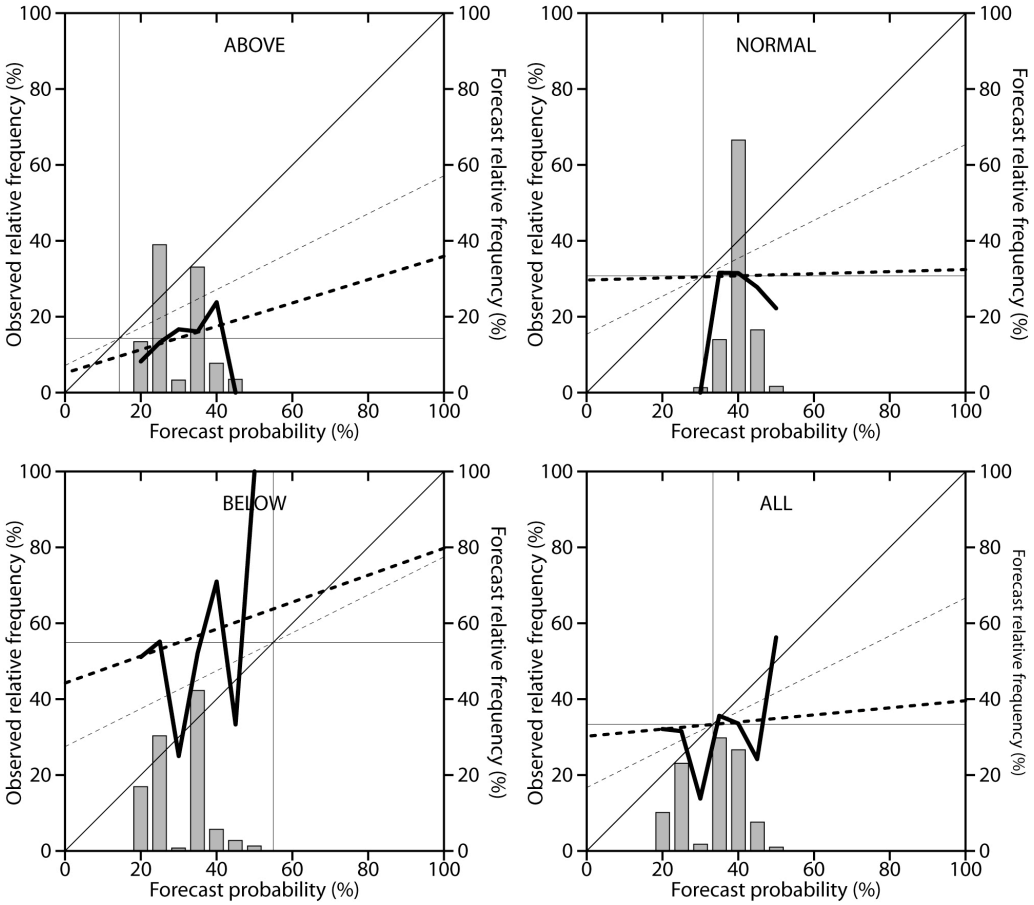


Figure 8. GHACOF attributes diagrams for the March to May season. The thick black line shows the reliability curve, and the thick dashed line is the least squares weighted regression fit to the reliability curve. The weights are shown by the grey bars, which indicate the relative frequency of forecasts in each 5% bin. The thin horizontal and vertical lines indicate the relative frequency of occurrence of rainfall in the respective category, while the thin diagonal represents the line of perfect reliability, and the thin dashed line the line of "no skill" as measured by the Brier score.

ii. September to December

The forecasts for the September to December period for the Greater Horn of Africa show much improved reliability and resolution compared to March to May (Figure 9). The above-normal category has particularly good reliability, increasing in frequency by about 10% for every 10% increase in the forecast probability. Below-normal rainfall increases by about 6%, and normal rainfall decreases by 5%. These results are mirrored by the measures of discrimination (Figure 6), with correct rates of 65% for above-normal, but of only 52% and 47% for below-normal and normal, respectively. The quality of the forecasts for the below-normal category were adversely affected by the forecasts with 50% probabilities, all of which were issued in 1999, when below-normal rainfall was not as widespread as implied. Because all these forecasts were issued in one year, the error bars on the reliability plot for this point should be considered large.

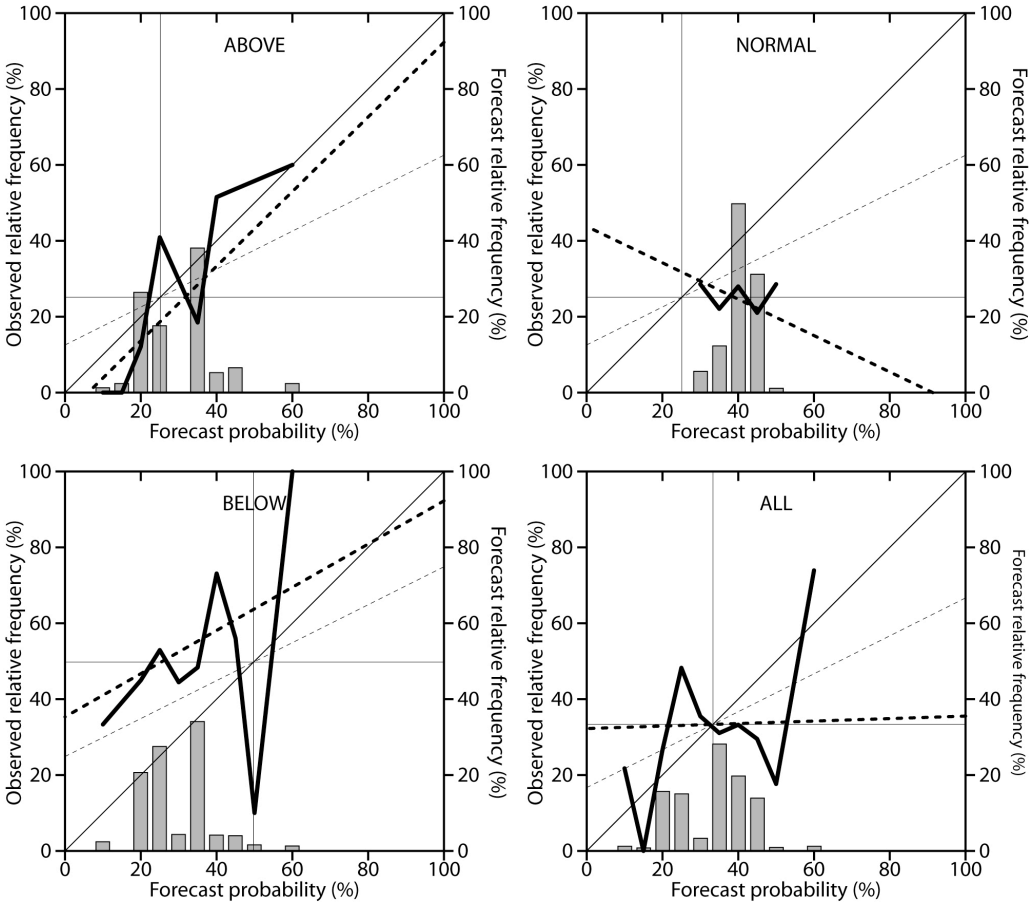


Figure 9. GHACOF attributes diagrams for the September to December season. The thick black line shows the reliability curve, and the thick dashed line is the least squares weighted regression fit to the reliability curve. The weights are shown by the grey bars, which indicate the relative frequency of forecasts in each 5% bin. The thin horizontal and vertical lines indicate the relative frequency of occurrence of rainfall in the respective category, while the thin diagonal represents the line of perfect reliability, and the thin dashed line the line of "no skill" as measured by the Brier score.

Despite the generally good resolution for the September - December forecasts, as with the March – May season there are some serious biases. Below-normal rainfall occurred 50% of the time, far more frequently than implied by the average forecast probability of 30%. In contrast, above-normal rainfall occurred less frequently than forecast (25% compared to 30%). The under-forecasting of the below-normal category was largely at the expense of serious over-forecasting of the normal category: the average forecast probability for this category was 40%, but normal rainfall occurred only 25% of the time. The degree of hedging is thus more severe than for March to May; very high probabilities (45% and higher) on the normal category were issued far more frequently in September to December, and probabilities were consistently higher than observed relative frequencies. The negative skill on this category together with the positive bias do not justify increasing the probability much above its climatological value.

The generally promising looking-results presented in Figures and 6 and 9 for the September – December forecasts, are brought into some question by the spatial distribution of the skill, which reveals an area of high skill in northern Sudan (Figure 10). This area is right at the end of its wet season when the forecast is made, and receives notable rainfall only in September. The seasonal forecast for northern Sudan is therefore more like an extended range weather forecast. However, these skill maps are subject to large sampling

errors, and so should not be over-interpreted. Nevertheless, the weak skill of the forecasts in areas further south is disappointing given the relatively high predictability (Mutai and Ward 2000).

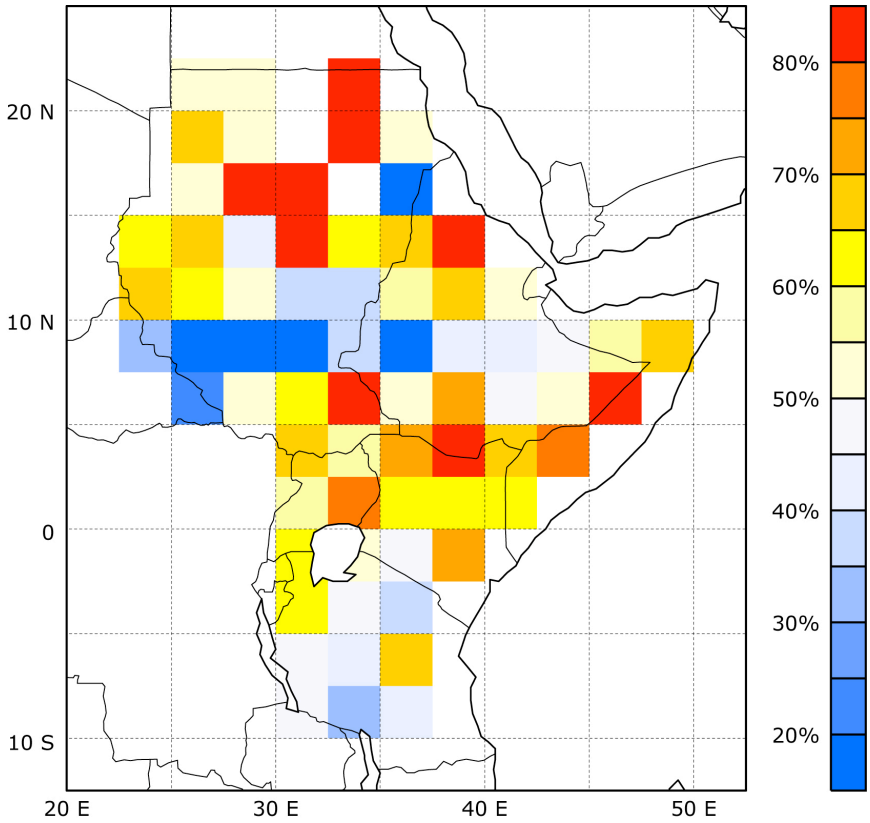


Figure 10. Map of generalized discrimination for the GHACOF September – December forecasts.

Overall the September – December forecasts for the Greater Horn of Africa, show weak skill and no resolution (the generalized discrimination is 52%, and a 10% increase in forecast probability implies no change in observed relative frequency). Despite the good results for the outer categories, overall results are poor because of the large biases in the forecasts. As for March to May, the biases are most severe for the below-normal category, with below-normal rainfall occurring far more frequently and extensively than forecast, and are affected by hedging on the normal category.

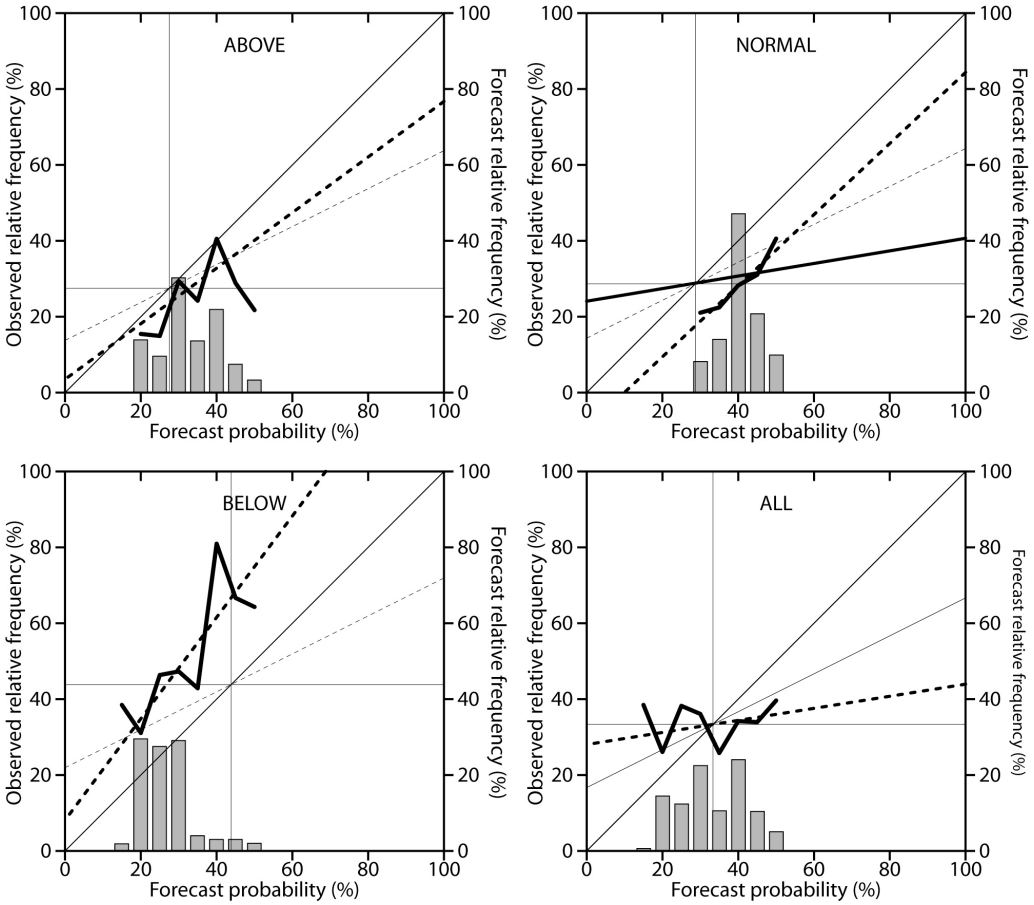


Figure 11. PRESAO attributes diagrams for the July to September season. The thick black line shows the reliability curve, and the thick dashed line is the least squares weighted regression fit to the reliability curve. The weights are shown by the grey bars, which indicate the relative frequency of forecasts in each 5% bin. The thin horizontal and vertical lines indicate the relative frequency of occurrence of rainfall in the respective category, while the thin diagonal represents the line of perfect reliability, and the thin dashed line the line of "no skill" as measured by the Brier score.

c. West Africa

The PRESAO forecasts for July to September show good reliability and resolution (Figure 11). The regression fit for the below-normal category suggests slight under-confidence (below-normal rainfall increases in frequency by 13% for every 10% increase in the forecast probability), but the fit is dominated by probabilities in only 3 bins (20%, 25%, and 30%), and there is large uncertainty over the reliability for other forecast values. In addition, the bias is reasonably strong: below-normal rainfall was observed 44% of the time, but the average forecast probability was 27%. Despite these weaknesses, the discriminatory power for this category is reasonably high (60%, Figure 12), in large part because of the skill from the high probabilities.

The bias is smaller for the above-normal category, and this category was over-forecast: it occurred 28% of the time over the verification period, compared to 33% suggested by the forecasts. However, the resolution for above-normal is weaker than for below-normal, with a 10% increase in forecast probability translating to a 7% increase in observed relative frequency. This weaker resolution is matched by a weaker, but reasonable discrimination (59%, Figure 12).

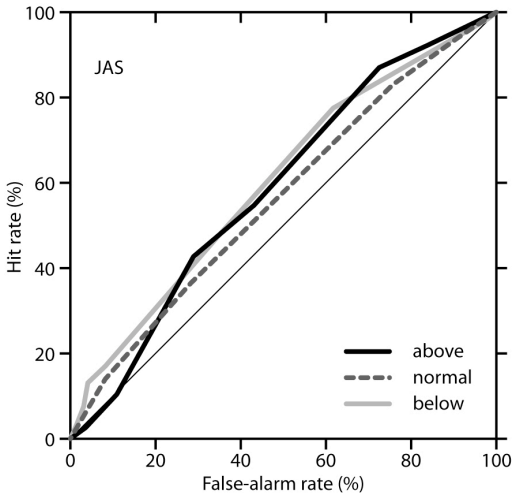


Figure 12. ROC diagram for the July to September (JAS) PRESAO forecasts. The thick black line is for above-normal rainfall, the dashed grey line for normal, and the light grey line for below-normal.

Unlike the other RCOFs, there is some skill for the normal category (although this skill vanished when alternative verification datasets were considered): it can be discriminated from other outcomes (56%, Figure 12), and there is a 9% increase in frequency for every 10% increase in forecast probability. However, there is a strong bias, with the category being over-forecast because of hedging: normal rainfall occurred 31% of the time, but the forecasts implied a 29% frequency.

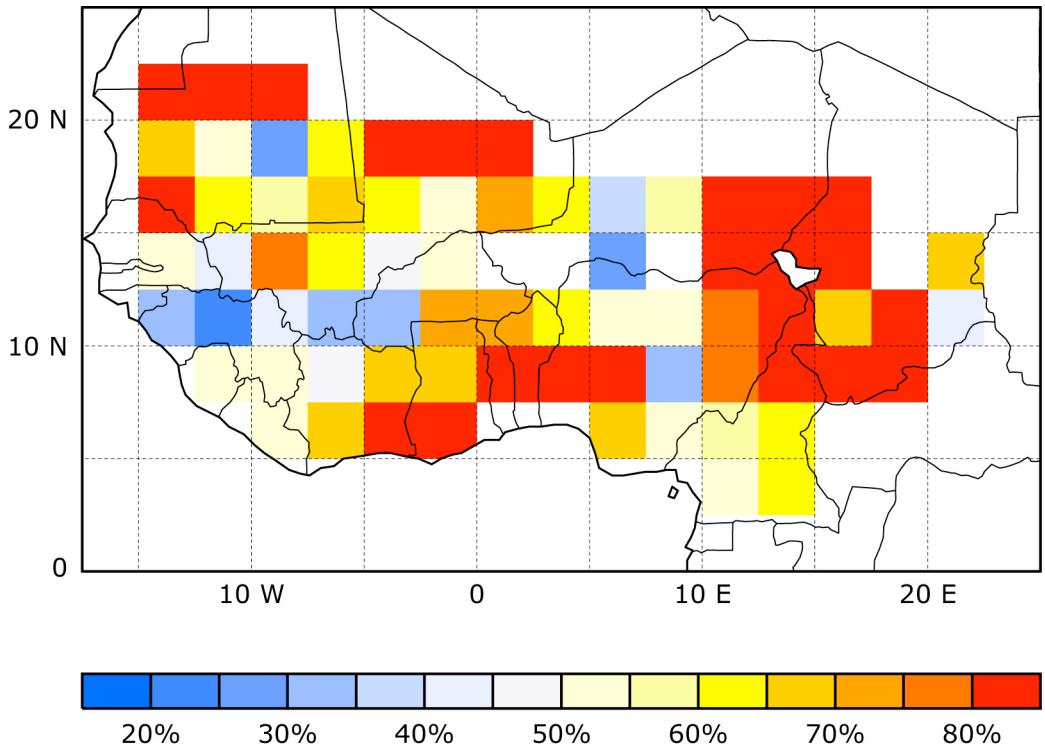


Figure 13. Map of generalized discrimination for the PRESAO July – September forecasts.

Figure 13 shows the spatial variation of generalized discrimination indicating the performance of the forecast system relative to climatology. The map indicates a more uniform distribution of positive skill than either of the other two regions. Most areas show positive skill, especially in the more northern, drier areas.

Overall the July to September forecasts for West Africa, show moderate skill and only weak resolution (the generalized discrimination is 60%, and a 10% increase in forecast probability implies only a 2% increase in observed relative frequency), despite the skill for the individual categories. The strong large biases (the forecasts did not provide indication of a shift towards below-normal rainfall over the verification period), and hedging negatively affect the overall results.

5. Conclusions and Recommendations

Ten years of GHACOF (Greater Horn of Africa) and PRESAO (West Africa) forecasts and ten to eleven years of SARCOF (Southern Africa) forecasts are verified using blended rain-gauge and satellite derived precipitation data from NOAA NCEP CPC AMS OPI and UEA CRU TS 2.1. For Southern Africa and the Greater Horn of Africa forecasts were verified for two seasons separately. The forecasts are consensus products based on statistical models, atmospheric circulation models from regional and international centres, and participants' expert interpretation.

All three regions indicate evidence of positive skill to varying degrees, thus fully endorsing the RCOF process, but they also show evidence of systematic errors, and in some cases the positive skill will not be immediately apparent to users, and thus there is considerable scope for improvement. The most ubiquitous error is the tendency to hedge the forecasts towards high probabilities on the normal category, presumably to avoid the risk of the forecasts being interpreted as being in error by two categories. This error reflects a tendency of the forecasters to continue to view and communicate the forecasts deterministically, which is most clearly evident in the way in which the previous season's forecast is verified using some form of hit score or other categorical scoring rule. It is imperative that an alternative, more suitable, verification procedure be implemented for reviewing the previous season's forecasts as a step towards encouraging the forecasters to forecast their true beliefs rather than their safest bets. This procedure would necessarily consider the probabilistic nature of the forecasts.

An effect of the hedging strategy is that normal rainfall was forecast to occur much more frequently and extensively than was observed. The tendency in some of the RCOFs is to issue relatively sharp probabilities on the normal category, with probabilities frequently reaching 50%. There is very little evidence of skill in forecasting increased probabilities of normal rainfall in any of the regions (with the possible exception of West Africa), and as an immediate correction it is recommended that probabilities for this category not exceed near-climatological values of about 35%, perhaps reaching 40% only in the rare cases with clear evidence of the normal category being the most likely outcome. Thus in all the RCOFs, where probabilities of 40% and higher on the normal category are issued frequently, consideration needs to be given towards reducing the bias towards forecasting normal rainfall.

Another effect that can be partly attributed to the strong hedging strategy is that in none of the cases where the verification period experienced climate conditions that were notably different from the climatological period were these trends clearly indicated by the forecasts. These trends were most marked in East Africa, with below-normal rainfall recorded at about 50% or more in the Greater Horn in both seasons, and thus a failure to indicate them has to be acknowledged as a major weakness of the RCOFs. It is recommended that the reasons

for these failures be investigated by re-forecasting the period using purely objective methods based on sea-surface temperature predictors and GCM outputs.

More generally, although in as few cases there is evidence of good reliability, there are clear difficulties in setting reliable probabilities for the forecasts. These difficulties, together with the strong evidence for hedging, make a strong case for implementing more objective methods for setting the forecasts than are currently used. Not only is there a strong need to reduce the subjective component of the process, but there is a need to ensure that reliable objective schemes are introduced. The contingency table approach, for example, that is used occasionally should be discouraged, despite its intuitive appeal, because of its very large sampling errors (Mason and Mimmack 2002). Methods based on error variance calculations and/or recalibrated ensemble approaches should be promoted.

Recognition should be given to the fact that in some regions there is evidence of good skill in forecasting only one of the two outer categories. Again the reasons for this result should be investigated, and genuine differences in skill for the alternative tendencies should be reflected in the forecasts. However, perhaps a more fundamental question is to address the reasons for the lack of skill. Although the RCOFs can take pride in the fact that they have produced some skilful forecasts over the last approximately 10 years, in many cases the skill is not immediately evident. In GHACOF, for example, the overall measures of skill for both seasons are very weak, and so it will not be immediately apparent to the users how to take any useful information from the forecasts. Here, but in the other regions too, a serious assessment of the methods used to produce the forecasts is required not only to improve the reliability and resolution of the forecasts, but also to work towards increasing their sharpness; for all three regions and all seasons the sharpness of the forecasts is low. Although there are a few cases of under-confidence, the forecasters are warned against being encouraged to simply issue slightly sharper forecasts based on the verification results, but instead to work towards the more objective forecasting procedures mentioned earlier.

References

Berri, G. J., P. L. Antico, and L. Goddard, 2005: Evaluation of the Climate Outlook Forums' seasonal precipitation forecasts of Southeast South America during 1998-2002. *Int. J. Climatol*. 25, 365-377.

Brier, G. W., and Allen, R. A. 1951: *Compendium of Meteorology*. *Am. Meteorol. Soc.*, pp. 841-884.

Janowiak, J. E., and P. Xie, 1999: CAMS-OPI: a global satellite-raingauge merged product for real-time precipitation monitoring applications. *J. Climate* 12, 3335-3342.

Jolliffe, I. T., and D. B. Stephenson, 2003. Introduction. In Jolliffe I. T., and D. B. Stephenson, Eds, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, Chichester, 1-12.

Mason, S. J., 2009: *Recommended Procedures for the Verification of Operational Seasonal Climate Forecasts*. WMO Technical Publication, under review.

Mason, S. J., and G. M. Mimmack, 2002: Comparison of some statistical methods of probabilistic forecasting of ENSO. *J. Climate* 15, 8–29.

Mason, S. J., and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.* 138, 331–349.

Mitchell, T. D., and P. D. Jones, 2005: An improved method of constructing a database of monthly climate observations and associated high resolution grids. *Int. J. Climatol.* 25, 693–712.

Mutai, C. C., and M. N. Ward, 2000: East African rainfall and the tropical circulation/convection on intraseasonal to interannual timescales. *J. Climate* 18, 3915–3939.

Ogallo L. J., P. Bessemoulin, J. P. Ceron, S. J. Mason, and S. J. Connor, 2008: Adapting to climate variability and change: the Climate Outlook Forum process. *J. World Meteor. Org.* (http://www.wmo.int/pages/publications/bulletin/ogallo_en.html).

Wilks, D. S., 2000: Diagnostic verification of Climate Prediction Center long lead outlooks, 1995–1998. *J. Climate* 13, 2389–2403.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego, 648 pp.

Wilks, D. S., and C. M. Godfrey, 2002: Diagnostic verification of the IRI Net Assessment forecasts, 1997–2000. *J. Climate* 15, 1369–1377.

Wilks, D. S., and A. H. Murphy, 1998: A case study of the use of statistical models in forecast verification: Precipitation probability forecasts. *Wea. Forecasting* 13, 795–810.